# Označevanje udeleženskih vlog v učnem korpusu za slovenščino

**Simon Krek,*+ Polona Gantar,♦ Kaja Dobrovoljc,† Iza Škrjanec‡**

* Laboratorij za umetno inteligenco, Institut »Jožef Stefan«, Jamova cesta 39, 1000 Ljubljana
+ Center za jezikovne vire in tehnologije Univerze v Ljubljani, Večna pot 113, 1000 Ljubljana
simon.krek@ijs.si
♦ Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana
apolonija.gantar@ff-uni.lj.si
† Zavod za uporabno slovenistiko, Trojina, Trg republike 3, 1000 Ljubljana
kaja.dobrovoljc@trojina.si
‡ Ljubljana
skrjanec.iza@gmail.com

**Povzetek**

V prispevku predstavimo postopek, nabor oznak, merila ter orodje za semantično označevanje učnega korpusa za slovenščino. V prvem delu prispevka predstavimo teoretična izhodišča raziskave in uporabljeno metodologijo, nato pa podrobno opišemo nabor oznak za semantično označevanje učnega korpusa za slovenščino in merila za njihovo določanje. Posebej izpostavimo konkurenčne udeleženske vloge in potencialne nove udeleženske vloge za razreševanje mejnih primerov. Prispevek zaključimo s kratkim povzetkom sprejetih odločitev in predvidenim nadaljnjim delom v okviru bilateralnega projekta Označevanje semantičnih vlog v slovenščini in hrvaščini.

**Semantic Role Labeling in the Training Corpus for Slovene**

The paper describes the procedure, tagset, criteria and tools for semantic role labeling in the training corpus for Slovene. In the first part we present the theoretical foundations of our research and the methodology. The following part includes a detailed description of the tagset used for semantic role labeling of Slovene, together with annotation criteria. Ambiguous cases are highlighted and potential now semantic roles are suggested for solving borderline cases. The paper finishes with a short summary of the decisions that were taken in the process, and future work in the context of the bilateral Slovene-Croatian project Semantic Role Labeling in Slovene and Croatian.

## 1 Uvod

Označevanje semantičnih vlog (ang. Semantic Role Labeling – SRL) je postopek, ki je z jezikoslovnega vidika namenjen (avtomatskemu) prepoznavanju udeleženskih vlog, z jezikovnotehnološkega pa razvoju sistemov za luščenje informacij, sistemov za odgovarjanje na vprašanja (ang. question answering system), izboljšavi delovanja skladenjskih razčlenjevalnikov ter strojnih prevajalnikov ipd. (Shen in Lapata, 2007; Christensen et al., 2011). Ker pomanjkanje konsenza glede različnih kategorij in meril za njihovo določanje, ki so danes sicer že na voljo za številne jezike, povzroča težave pri čezjezikovnem modelu semantičnega označevanja, mora po našem mnenju uspešen sistem meril in oznak za označevanje udeleženskih vlog ali natančneje predikatno-argumentnih razmerij (a) zagotavljati nabor kategorij, ki je kar najbolj optimalen, tj. pokriti vse (v našem primeru za slovenščino) ključne udeleženske vloge in hkrati (b) ne vsebovati kategorij, ki so prepodrobne ali medsebojno prekrivne, (c) temeljiti primarno na semantičnih in ne na morfoloških, leksikalnih ali skladenjskih lastnostih, (d) omogočati formalni opis oz. uporabnost v jezikovnotehnoloških aplikacijah ter (e) biti čim bolj kompatibilen s kategorijami in merili, ki veljajo za druge jezike (prim. Petukhova in Bunt, 2008: 39). V ta namen je bil v okviru projekta izdelave učnega korpusa za označevanje semantičnih vlog za slovenščino izdelan sistem meril za prepoznavanje in označevanje udeleženskih vlog za slovenščino. Naš cilj je bil ročno označiti polovico skladenjsko označenega dela učnega korpusa ssj500k,[1] na njegovi podlagi pa naj bi bilo v prihodnje mogoče avtomatsko označiti tudi obsežnejše korpuse.

V nadaljevanju prispevka predstavimo izhodišča za določitev semantičnih kategorij ter nabor oznak za slovenščino, postopek označevanja in orodje za semantično označevanje učnega korpusa za slovenščino.

## 2 Teoretično in metodološko ozadje

Pri izbiri metode semantičnega označevanja in določanju semantičnih kategorij za slovenščino smo najprej analizirali posamezne pristope, ki so bili razviti in uporabljeni za druge jezike, npr. PropBank (Palmer et al., 2005), Verbnet (Kipper et al., 2006) in FrameNet (Backer et al., 1998) za angleščino, AnCora (Taulé et al., 2011) za španščino, SoNaR (Schuurman et al., 2010) za nizozemščino. Poleg tega pa še nabor oznak za hrvaščino (Filko et al., 2012) in češki valenčni leksikon Vallex.[2] Osredotočili smo se na primerjavo formalnih opisov (tj. naborov semantičnih oznak) za posamezne udeleženske vloge ter meril za njihovo določanje. Z vidika optimizacije nabora oznak, ki bi zagotavljal dovolj robusten sistem in hkrati v čim večji meri upošteval specifike slovenščine, smo upoštevali še stopnjo semantične razdrobljenosti, ki jo predvideva posamezni sistem, in dejstvo, da za slovenščino nimamo na voljo strojno berljivega leksikona glagolske vezljivosti. Poleg tega smo merila za semantično označevanje želeli določiti tako, da bodo

[1] Opis in prenos korpusa: http://www.slovenscina.eu/tehnologije/ucni-korpus.

[2] Vallex: http://ufal.mff.cuni.cz/vallex.